DOCUMENT RESUME

ED 269 455 TM 860 306

AUTHOR Cantor, Nancy K.; Hoover, H. D.
TITLE The Reliability and Validity of Writing Assessment: An Investigation of Rater, Prompt within Mode, and Prompt between Mode Sources of Error.
PUB DATE Apr 86
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Cues; Elementary Secondary Education; *Error of Measurement; *Essay Tests; Interrater Reliability; Language Proficiency; *Scores; Standardized Tests; Tables (Data); Test Construction; *Test Reliability; *Test Validity; *Writing Evaluation
IDENTIFIERS Cognitive Abilities Test; Iowa Tests of Basic Skills; *Iowa Tests of Basic Skills Writing Supplement

ABSTRACT

This paper isolates and examines separately three distinct sources of error in essay scores: lack of agreement between raters; inconsistencies in performance within mode of discourse, and inconsistencies in performance between modes of discourse. Essay prompts in the Iowa Tests of Basic Skills (ITBS) Writing Supplement were designed to assess students' skills in five modes of discourse: the narrative, the explanation, the description, the informative report, and the persuasive essay. A subsample of 46,000 students participated in the spring, 1985, standardization of the ITBS. One quarter of these 46,000 were assigned to the standardization of the ITBS Writing Supplement, and this report considers data from 10 of 20 subsamples from this quarter. Estimates of reader reliability range from .83 to .95, those for within-mode score reliability range from .37 to .69, and those for between-mode score reliability are generally in the .30s. This study reports several indicators of test validity: (1) measures of prompt appropriateness and scoring protocol usefulness; (2) evidence that the essays collected measure both small increments of growth and significant increases in writing proficiency; and (3) evidence for the discriminant validity of the essays. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED269455

The Reliability and Validity of Writing Assessment: An Investigation of Rater, Prompt within Mode, and Prompt between Mode Sources of Error

Nancy K. Cantor and H. D. Hoover
The University of Iowa

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
N. K. Cantor

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper presented at the
Annual Meeting of the
American Educational Research Association
San Francisco
April, 1986

TM 860 306

ABSTRACT

This paper isolates and considers separately three distinct sources of error in essay scores: lack of agreement among scorers; inconsistencies in performance within mode of discourse; and inconsistencies in performance between modes of discourse. Estimates of reader reliability range from .83 to .95, those for within-mode score reliability range from .37 to .69, and those for between-mode score reliability are generally in the .30s.

In addition, the paper reports several indications of test validity: measures of prompt appropriateness and scoring protocol usefulness; evidence that the essays collected measure both small increments of growth and significant increases in writing proficiency; and evidence for the discriminant validity of the essays.

INTRODUCTION

The problems inherent in measuring writing ability have created a dilemma at least since the turn of the century, when the College Board offered its first tests to the college-bound public. For the first fifty or sixty years of this century, measurement experts were most strongly influenced by three major limitations in the direct assessment of writing: time considerations generally preclude the collection of more than one essay per student, and the scoring of that essay is inevitably less consistent, and more expensive, than the scoring of objective items. During the same time period, language arts and rhetoric programs were generally dominated by the formalist tradition, with its emphasis on the language of the text. The consequence of these two parallel forces was that large-scale testing programs, although they may have experimented with the collection of writing samples, generally limited themselves in practice to the assessment of students' knowledge of standard written English via the objective test format.

Well-constructed objective tests provide highly reliable measurement of the skills they are designed to measure. By their nature, however, they tap a student's editorial rather than productive skills. In addition, they generally emphasize sentence-level rather than discourse-level skills.

Within the last twenty years, language arts and rhetoric programs

---

Many people were involved in the collection and analysis of these data. We are particularly grateful to Judy Hu, Jim Keasler, Carole Meinders, Sharon Mullin, Maxine Strege, the staff of Measurement, Inc., and the staff and students of the participating school systems.

have moved, sometimes gradually, sometimes explosively, toward an emphasis on productive and discourse-level skills (Connors, Ede, & Lunsford, 1984; Donovan & McClelland, 1980; Freedman & Pringle, 1980). Increasingly, good writing is defined as the ability to generate ideas of substance as well as the ability to express those ideas effectively. Both theorists and practitioners are finding it increasingly difficult (and increasingly artificial) to talk about good writing as distinct from good thinking. Lee Odell undoubtedly expresses the views of many of his colleagues when he defines competence in writing as "the ability to discover what one wishes to say and to convey one's message through language, syntax, and content that are appropriate to one's audience and purpose" (Odell, 1981, p. 103). The current stance of most experts in language arts and rhetoric vis-à-vis measurement seems to be: if direct measures of writing allow students to demonstrate skills that are of interest to us and that cannot be measured through objective items, then those measures are desirable, and their lower reliability can be tolerated.

At the same time, measurement specialists are demonstrating more general interest in the problem of the writing sample. Twenty or more correlational and experimental studies published since 1966 specifically address the reliability and validity of the essay score.[1] The purposes,

[1]See Godshalk, Swineford, & Coffman, 1966; Coffman, 1966; Myers, McConville, & Coffman, 1966; Follman & Anderson, 1967; Coffman, 1971; Veal & Tillman, 1971; Breland & Gaynor, 1979; Crocker, Ondrasik, & Lamme, 1979; Huntley, Schmeiser, & Stiggins, 1979; Powills, Bowers, & Conlan, 1979; Hogan & Mishler, 1980; Breland & Griswold, 1981; Culpepper & Ramsdell, 1982; Moss, Cole, & Khampalikit, 1982; Quellmalz, Capell, & Chou, 1982; Prater & Padia, 1983; Veal & Hudson, 1983; Weiss & Jackson, 1983; Arrasmith, Sheehan, & Applebaum, 1984; DeShields, Hsieh, & Frost, 1984; Carlson, Bridgeman, Camp, & Waanders, 1985.

and hence the designs, of these investigations vary widely, as do their results. Sample sizes range from 5 (DeShields et al., 1984) to over 10,000 (Breland & Griswold, 1981). Interrater reliabilities range from .366 (Godshalk et al, 1966) to .92 (Hogan & Mishler, 1980). Estimates of score reliability (correlations between scores on two or more essays) range from .221 (Godshalk et al., 1966) to .81 (Hogan & Mishler, 1980). Correlations between essay scores and scores on objective measures of language skills range from .20 (Moss et al., 1982) to .81 (Carlson et al., 1985).

Informed by these empirical investigations and by theoretical articles on prompt development (Ruth & Murphy, 1984) and on scoring techniques (Cooper & Odell, 1977; White, 1984), we are now able to develop direct measures of writing with a fair degree of confidence in their reliability. And we do. By the spring of 1985, it was reported that 33 states have or are in the process of developing statewide writing assessment programs (Afflerbach, 1985).

However, essay measures, like their objective counterparts, cannot be assumed to be uniformly reliable. And we do not yet, with any degree of consistency, apply to our large-scale instruments the techniques for estimating reliability and validity which we have learned from our research studies. Too often we do not report these estimates at all. Without this information, the test user has no way of judging the quality of the instrument he or she chooses or is assigned.

This brings us to the purpose of this report: to describe a standardization design which is both practical and informative. By this I mean a design which is reasonable in the demands it makes on

subjects and scorers, thereby permitting the national standardization of an essay instrument, but one which simultaneously provides data on which we can base estimates of reliability and validity.

METHOD

Instruments

Essay prompts in the *Iowa Tests of Basic Skills Writing Supplement* were designed to assess students' skills in five modes of discourse: the narrative, the explanation, the description, the informative report, and the persuasive essay. For each of these modes, two parallel prompts were developed for students in Grades 3 through 5. A second pair of prompts in each mode was developed for students in Grades 6 through 8. All twenty prompts were designed to be modal (representative of the most salient features of the mode of discourse), explicit (students were reminded of the unique demands of the mode of discourse represented), and fair (capable of eliciting unbiased responses from students of both sexes and a variety of racial and ethnic groups in both urban and rural school systems in all regions of the country).

Objective instruments used in this study were the *Iowa Tests of Basic Skills (ITBS)* and the *Cognitive Abilities Test (CogAT)*. The regular battery of the *ITBS* measures achievement in vocabulary, reading, language (four tests: spelling, capitalization, punctuation, and usage and expression), work-study skills (two tests: visual materials and reference materials), and mathematics (three tests: concepts, problem-solving, and computation). The *CogAT* measures students' aptitudes in working with three basic types of symbols: verbal (three tests: verbal

classification, sentence completion, and verbal analogies); quantitative (three tests: quantitative relations, number series, and equation building); and geometric or spatial (three tests: figure classification, figure analogies, and figure analysis).

Both the objective *ITBS* tests and the essay prompts were administered within a two-week period in April and May of 1985. *CogAT* scores were collected the previous fall.

Sample

The main standardization of the *ITBS* occurred in the fall of 1984. A subsample of approximately 46,000 students participated in the spring, 1985, standardization of the *ITBS*. The sample was selected to be representative of the national school population with respect to ability and achievement. Public school sampling units were selected on the basis of region of the country (4 areas), school district size (9 levels), and socioeconomic status (5 levels). Parochial schools were stratified by geographic region and the enrollment of the diocesan school system to which they belonged. Stratification variables for private, non-Catholic schools were geographic region and type of school.

Of these 46,000 students, one quarter were assigned to the standardization of the *ITBS Writing Supplement*. Participating students were administered the 11 subtests of the regular *ITBS* battery and 2 supplementary essays. Approximately half the students were assigned parallel prompts within the same mode of discourse; the other half responded to essay prompts in different modes of discourse. Students who failed to write both essays or who failed to code their essays for administration order were dropped from the sample. Approximately 10,000 students, representing 55 buildings in 24 states, returned coded pairs of essays.

These 10,000 students were then divided into 20 separate samples, each determined by the essay prompt to which the student responded first. In order to ensure the representativeness of individual samples, each was stratified on the basis of the student's language total score from the regular battery of the *ITBS*. Whenever possible, 10 students were selected from each decile of the language total distribution for their grade level. In most cases, this resulted in the selection of 100 papers per grade or 300 papers per prompt.

This report considers data from 10 of these 20 subsamples, those in which the student's first essay was written in response to one of the 10 essay prompts which will make up Form G of the *ITBS Writing Supplement*. The data consist of a pair of essay scores for each of 2835 students, *ITBS* scores for 2557 of those students, and *CogAT* scores for 2593 of them.

Essay Scoring

Essays were scored on a 4-point, focused holistic score scale by experienced raters employed by a commercial scoring firm. Papers which were blank, totally illegible, written in a language other than English, or written on a topic other than the one assigned were designated unscorable and assigned a value of zero in all numerical analyses.

A separate set of training and scoring materials was prepared for each prompt. Each set contained a mode-specific scoring protocol, three anchor papers at each score point, and 30 training papers.

Scoring protocols provided an abstract discussion of the criteria for each point in the score scale. They were designed to define the unique demands of each mode of discourse, as well as the general demand across modes for fluency, generativity, and the presence of a personal

voice. Emphasis generally was on thoughtful content, logical organization, and an original and appealing style. Language mechanics played a major role only in those instances where they so limited the student that an adequate judgment could not be made about the essay's content, organization, or style.

Anchor papers provided concrete examples of the criteria set forth in the scoring protocol. Training papers were used to bring the raters into conformity with the standards established by the scoring protocol and the anchor papers.

Each essay was scored by two independent raters. In those cases where the two scores differed by more than one point, the paper was scored by a third rater. The final score assigned each essay was the sum of its two closest ratings.

RESULTS AND DISCUSSION

Reliability

Three different types of reliability coefficients, indicating different sources of error in the ratings assigned to individual pupil essays, are of interest in the context of the present study. Correlation coefficients provide estimates of each type. The first, interrater reliability, is estimated by the correlations between the two independent ratings given each essay. Intraclass correlations, calculated separately for each grade and mode of discourse and then averaged across grades within mode, are presented in Table 1. They range from .78 for Grade 3 informative reports to .98 for Grade 4 persuasive essays.

Average interrater reliabilities for narrative and descriptive essays and for informative reports (.88, .88, and .83, respectively) are in the upper end of the range reported in research studies which use essay scales requiring qualitative judgments about content, language, and style. Interrater reliabilities for explanatory and persuasive essays (both .95) are higher than those previously reported.

The second and third types of reliability estimates take into account variability within and between modes of discourse across time and across topic. These correlations, again calculated separately for grades and modes and averaged across grades within mode, are presented in Table 2.

Within-mode reliabilities are estimated by the correlations between scores on two essays written on different occasions and on different topics, where the assigned topics are intended to elicit responses in the same mode of discourse. They are analogous to parallel-forms reliabilities.

Within-mode correlations fall into three distinct groups: Narrative reliabilities, which average .69, are quite high relative to essay-essay correlations generally reported. Those for explanation (average of .50) are moderate. Those for descriptive and persuasive essays and for informative reports (.38, .37, and .40), although well within previously-reported ranges, are less than satisfactory for reporting individual pupil scores.

The variability of these correlations from mode to mode adds quantitative evidence to a premise we had developed earlier, based on a review of curriculum materials prior to prompt development and on the

impressionistic reading of hundreds of essays in the course of developing scoring protocols and selecting illustrative anchor papers. Students in the upper elementary and middle school grades seem to receive most of their writing instruction and most of their writing practice in the narrative mode. They are most commonly asked to "write a story," and they have developed personal algorithms for doing so. The quality of these algorithms varies substantially, but they tend to be applied consistently from day to day and from topic to topic.

One measure of the degree to which narrative formulas are overlearned and procedures for other modes neglected is the frequency with which students assigned descriptive, explanatory, and persuasive topics tend to lapse into the more familiar narrative mode when they write themselves into a corner in the assigned mode. These lapses lead to essays judged inappropriate to audience and purpose.

It seems reasonable to expect, and the data suggest, that this inclination to narrative varies considerably from topic to topic. In the explanatory mode, which is relatively straightforward, performance across topic is more consistent. The precepts for descriptive and persuasive writing are conceptually more difficult, and performance in these modes is less consistent across topics.

Preliminary readings of informative reports suggest that the lack of stability in these scores is probably due to our insistence that students receiving scores of "3" or "4" in this mode generate something recognizable as a thesis and organize their reports around it, using only the source material relevant to what they considered "the most

important or most interesting main idea." This is a difficult and frequently untaught cognitive process, and our experience suggests that it is applied inconsistently across topics.

Between-mode reliabilities are estimated by correlations between scores on two essays written on different occasions and on different topics, where the assigned topics are intended to elicit responses in different modes of discourse. They may be thought of as non-parallel forms reliabilities or as correlations between divergent components of a general construct. In this sense, they are analogous to correlations between reading comprehension scores obtained from readings of contemporary poetry and those obtained from readings of a general science text.

Between-mode correlations range from .32 for persuasive topics paired with topics in other modes to .40 for descriptive topics paired with topics in other modes. Correlations this modest generally add support to the growing awareness that writing is not a unidimensional trait, that we cannot assume that "a good writer is a good writer" (Quellmalz et al., 1982), and that it is presumptuous to base an <u>overall</u> estimate of writing ability on the collection of a single essay per student.

Table 3 summarizes the reliability data presented in Tables 1 and 2 and provides averages of interrater, within-mode, and between-mode reliabilities across modes of discourse.

Validity

In the absence of an overwhelmingly persuasive criterion measure, the test developer generally approaches questions of test validity

from a variety of angles, hoping that in the aggregate they will provide the information the user needs. Some efforts at maximizing validity were alluded to in earlier sections of this report: the differentiation between modes of discourse; the attempt to develop modal, explicit, fair prompts; and the development of scoring protocols with an emphasis on the appropriateness of language, syntax, and content to audience and purpose. This section reports the results of quantitative efforts to measure validity.

A major concern of the test user, given standardized essay prompts (those allowing no choice on the student's part), is the universal appropriateness of the prompts. The proportion of blank papers or papers written on a topic other than the one assigned provides an estimate of prompt appropriateness.

A parallel concern is the usefulness of the scoring protocols in distinguishing score points. The proportion of papers on which two raters, using the same protocol, differ by at least one point provides an estimate of the usefulness of the protocols.

The data in Table 4 address these two concerns. The proportion of unscorable papers (which includes illegible and non-English, as well as blank and off-topic papers) represents an upper limit on the proportion of students in a given grade for whom the assigned topic was inappropriate. With the exception of persuasive essays in Grade 3, proportions of unscorable papers are consistently below 3%, which seems to us an extremely acceptable level of response. Only well-above-average schools should be advised to assign persuasive essays in Grade 3.

All papers were scored by two raters. Provision was made for a

third rater when the two scores differed by more than one point. In fact, this degree of discrepancy occurred only once in the scoring of the 5670 essays. Discrepant-score papers, therefore, are defined as those in which the original two scores differ by precisely one point. With the exception of the informative report, proportions of discrepant-score papers in all modes are within acceptable limits for large-scale scoring efforts. Frequency distributions for report scores indicate that the high proportion of discrepant-score papers in this mode results from inconsistencies in assigning "1" and "2" scores to these papers. This is the point at which reports which essentially reproduce the source material should be distinguished from those which reorganize it; it is clearly a point which should be emphasized and elaborated in future scoring sessions.

A third validity concern is based on the assumption that writing performance improves over time. Because empirical, pretest-posttest investigations of one instructional method versus another fail so frequently to demonstrate significant growth by either method (Hillocks, 1985), we must assume that this growth is very slow indeed. This suggests that valid writing assessment must somehow unite the narrow score range common to most essay scales with extreme sensitivity to small increments of growth.

The data displayed in Tables 5 and 6 suggest that the *Writing Supplement* is sensitive to improved performance. Table 5 contains means of essay scores by grade within mode. Because students in Grades 3 through 5 responded to the same prompts and because their essays were scored in accordance with common protocols by raters unaware of the

students' grade levels, increases in mean scores across these grades provide evidence for this sensitivity. The same is true for increases in mean scores across Grades 6 through 8. Mean scores between Grades 5 and 6 should not be compared.

In all cases, mean scores do increase across grades. In some modes, growth seems fairly consistent. In others, it appears to accelerate and plateau from grade to grade. Without confirming theoretical or empirical evidence, there is no way to judge whether this is truly the case or whether it is an artifact of the test. In no case, however, do mean scores remain constant or decline from grade to grade.

Table 6 contains the results of tests of significance between Grade 3 and Grade 5 means and between Grade 6 and Grade 8 means. These results suggest that not only is the test sensitive to small increments of growth, it is also capable of revealing significant differences in performance levels over large numbers of subjects and a two-year time span.

A fourth concern of test users is lack of redundancy or discriminant validity. In particular, essay samples are useful only to the extent that they measure a construct which cannot be measured more efficiently by objective tests of language skills or verbal aptitude. The correlations between essay scores and objective scores in Tables 7 through 11 address this concern. Although sampling error results in substantial fluctuations in particular grades and particular modes, the summary of averages provided in Table 12 suggests that correlations generally range from the upper .30s to the low .50s for individual subtests and in the .40s and .50s for total or composite scores.

Table 13 displays disattenuated values of the average correlations between essay scores and *ITBS* language total scores and between essay scores and *CogAT* verbal scores. These values represent estimates of the correlations that would be expected if each test were perfectly reliable. Disattenuated correlations for the essay scores were calculated using average within-grade within-mode reliabilities from Tables 7 through 11. Those for objective scores were calculated using average within-grade K-R 20 reliabilities obtained in the fall, 1985, national standardizations of the *ITBS* and *CogAT*. To the extent that disattenuated correlations are smaller than 1.00, they indicate that the *Writing Supplement* is assessing aspects of language ability not already assessed by these objective measures.

SUMMARY

This paper has isolated and examined separately three distinct sources of error in essay scores: lack of agreement between raters; inconsistencies in performance within mode of discourse; and inconsistencies in performance between modes of discourse. Estimates of reader reliability are occasionally reported in the manuals of large-scale assessment instruments. Estimates of score reliability, though common in research studies, are rare in technical manuals. Our results indicate that estimates of these three types of error differ substantially in magnitude and that reader reliability alone should not be used as an estimate of test reliability.

Estimates of interrater reliability range from .83 to .95; these are well within or above the range commonly reported in research studies.

Estimates of within-mode score reliability range from .37 to .69. With a proviso on testing only those students who have been instructed in a given mode of discourse, these reliabilities would be regarded as well within the acceptable range by those educators with an interest in essay sampling. Estimates of between-mode score reliability are in the .30s and generally confirm the growing reluctance to interpret scores in one mode of discourse as evidence of overall writing proficiency.

In addition, this study reports several indications of test validity, some of which have not been reported elsewhere. Essay prompts and scoring protocols, as measured by proportions of unscorable and discrepant-score papers, are generally appropriate and useful. An examination of differences in grade-to-grade means and the performance of tests of significance suggest that the essays collected measure both small increments of growth and significant gains in writing performance. Finally, correlations between essay scores and objective measures of aptitude and achievement, including disattenuated values, suggest that these essays measure aspects of language ability not measured by these objective tests.

REFERENCES

Afflerbach, P. (1986). Statewide assessment of writing. (TME Report 87). Princeton, NJ: Educational Testing Service.

Arrasmith, D. G., Sheehan, D. S., & Applebaum, W. R. (1984). A comparison of the selected-response strategy and the constructed-response strategy for assessment of a third-grade writing task. Journal of Educational Research, 77, 172-177.

Breland, H. M. & Gaynor, J. L. (1979). A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 16, 119-128.

Breland, H. M. & Griswold, P. A. (1981). Group comparisons for basic skills measures. (College Board Report No. 81-6; ETS RR No. 81-21). New York: College Entrance Examination Board.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English. (TOEFL Research Report 19; ETS RR 85-21). Princeton, NJ: Educational Testing Service.

Coffman, W. E. (1966). On the validity of essay tests of achievement. Journal of Educational Measurement, 3, 151-156.

Coffman, W. E. (1971). On the reliability of ratings of essay examinations in English. Research in the Teaching of English, 5, 24-36.

Connors, R. J., Ede, L. S., & Lunsford, A. A. (Eds). (1984). Essays on Classical Rhetoric and Modern Discourse. Carbondale, IL: Southern Illinois University Press.

Cooper, C. R. & Odell, L. (Eds.). (1977). Evaluating Writing: Describing, Measuring, Judging. Urbana, IL: National Council of Teachers of English.

Crocker, L., Ondrasik, T., & Lamme, L. (1979). Predicting children's writing performance from standardized achievement tests. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED 174 684).

Culpepper, M. M. & Ramsdell, R. (1982). A comparison of a multiple choice and an essay test of writing skills. Research in the Teaching of English, 16, 295-297.

DeShields, S. M., Hsieh, H. K., & Frost, D. K. (1984). The measurement of writing skills: some problems and a probable solution. Educational and Psychological Measurement, 44, 101-112.

Donovan, T. R. & McClelland, B. W. (Eds.). (1980). Eight Approaches to Teaching Composition. Urbana, IL: National Council of Teachers of English.

Follman, J. C. & Anderson, J. A. (1967). An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1, 190-200.

Freedman, A. & Pringle, I. (Eds.). (1980). Reinventing the Rhetorical Tradition. Ottawa: Canadian Council of Teachers of English.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). The Measurement of Writing Ability. New York: College Entrance Examination Board.

Hillocks, G., Jr. (1986). Research on Written Composition: New Directions for Teaching. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills and National Conference on Research in English.

Hogan, T. P. & Mishler, C. (1980). Relationships between essay tests and objective tests of language skills for elementary school students. Journal of Educational Measurement, 17, 219-227.

Huntley, R. M., Schmeiser, C. B., & Stiggins, R. J. (1979). The assessment of rhetorical proficiency: The role of objective tests and writing samples. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED 173 419).

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at Grades 4, 7, and 10. Journal of Educational Measurement, 19, 37-47.

Myers, A. E., McConville, C. B., & Coffman, W. E. (1966). Simplex structure in the grading of essay tests. Educational and Psychological Measurement, 26, 41-54.

Odell, L. (1981). Defining and assessing competence in writing. In C. R. Cooper (Ed.), The Nature and Measurement of Competency in English (pp. 95-138). Urbana, IL: National Council of Teachers of English.

Powills, J. A., Bowers, R., & Conlan, G. (1979). Holistic essay scoring: An application of the model for the evaluation of writing ability and the measurement of growth in writing ability over time. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 174 679).

Prater, D. & Padia, W. (1983). Effects of modes of discourse on writing performance in Grades Four and Six. Research in the Teaching of English, 17, 127-134.

Quellmalz, E. S., Capell, F. J., & Chou, C-P. (1982). Effects of discourse and response mode on the measurement of writing competence. Journal of Educational Measurement, 19, 241-258.

Ruth, L. & Murphy, S. (1984). Designing topics for writing assessment: Problems of meaning. College Composition and Communication, 35, 410-422.

Veal, L. R. & Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. Research in the Teaching of English, 17, 290-296.

Veal, L. R. & Tillman, M. (1971). Mode of discourse variation in the evaluation of children's writing. Research in the Teaching of English, 5, 37-45.

Weiss, D. & Jackson, R. (1983). The validity of the Descriptive Tests of Language Skills: Relationships to direct measures of writing ability and to grades in introductory college English courses. (College Board Report No. 83-4; ETS RR No. 83-27). New York: College Entrance Examination Board.

White, E. M. (1984). Holisticism. College Composition and Communication, 35, 400-409.

Table 1
Interrater Reliabilities

Mode of Discourse	Grade 3		4		5		6		7		8		Average Within-Grade Reliability	
	N	r	N	r	N	r	N	r	N	r	N	r	N	r
Narrative	100	.82	112	.87	100	.91	103	.83	101	.90	104	.90	620	.88
Explanation	66	.97	101	.94	100	.92	100	.98	67	.90	95	.94	529	.95
Description	100	.85	100	.80	100	.86	100	.88	98	.93	42	.94	540	.88
Informative Report	101	.78	100	.89	100	.80	100	.84	72	.85	100	.79	573	.83
Persuasive Essay	100	.94	101	.98	72	.96	100	.96	101	.91	99	.88	573	.95

Within-grade interrater reliabilities are intraclass correlations between the two ratings given each essay. The use of intraclass correlations eliminates inflated values due to systematic leniency or harshness of scoring on the part of individual raters.

Average within-grade reliabilities in this and all subsequent tables were calculated from Fisher z-transformations using weighted within-grade N-counts.

Table 2
Between Prompt Reliabilities

Mode of Discourse	Grade 3		4		5		6		7		8		Average Within-Grade Reliability	
	N	r	N	r	N	r	N	r	N	r	N	r	N	r
Narrative														
Within mode	58	.66	40	.68	47	.61	37	.67	57	.77	40	.71	279	.69
Between mode	91	.39	118	.37	104	.42	114	.32	130	.31	80	.21	586	.34
Explanation														
Within mode	60	.45	40	.54	40	.47	66	.57	12	-.47	.64	.61	282	.50
Between mode	59	.16	122	.40	109	.37	152	.45	146	.48	148	.22	736	.37
Description														
Within mode	60	.57	68	.24	46	.13	40	.40	51	.40	.19	.62	284	.38
Between mode	87	.43	165	.45	167	.42	94	.33	98	.35	.53	.39	664	.40
Informative Report														
Within mode	47	.12	39	.39	51	.45	48	.62	24	.18	46	.46	255	.40
Between mode	54	.17	61	.36	49	.32	112	.29	91	.42	76	.36	443	.33
Persuasive Essay														
Within mode	51	.32	54	.39	72	.27	52	.47	11	.71	82	.33	322	.37
Between mode	89	.47	78	.25	54	.38	48	.18	87	.25	17	.38	373	.32

Within mode pairs are those in which both essays are in the designated mode.

Between mode reliabilities are the weighted average of the correlations from each subsample in which an essay in the designated mode was paired with an essay in another mode.

Table 3
A Summary of Interrater and Between Prompt Reliabilities
Averaged across Grades 3-8

Mode of Discourse	Interrater Reliabilities		Between Prompt Reliabilities	
			Within Mode	Between Mode
Narrative	.88	(.94)	.69	.34
Explanation	.95	(.97)	.50	.37
Description	.88	(.94)	.38	.40
Informative Report	.83	(.91)	.40	.33
Persuasive Essay	.95	(.97)	.37	.32
Average	.91	(.95)	.48	.36

Parenthetical interrater reliabilities have been corrected by the Spearman-Brown formula to reflect the reliability of the total score (rating 1 + rating 2).

Table 4
Percents of Unscorable and Discrepant Score Papers

Mode of Discourse	Grade											
	3		4		5		6		7		8	
	U	DS	U	DS	U	DS	U	DS	U	DS	U	DS
Narrative	3	18	0	22	1	12	0	25	1	17	2	16
Explanation	2	4	3	10	0	13	0	4	0	15	8	15
Description	1	8	0	14	0	17	0	14	2	8	0	14
Informative Report	0	23	0	22	0	27	3	22	0	26	0	31
Persuasive Essay	11	12	3	3	1	4	0	11	0	16	1	15

Unscorable (U) papers are blank, totally illegible, written in a language other than English, or written on a topic other than the one assigned.

Discrepant score (DS) papers are those on which Score 1 and Score 2 differed by one point.

23

Table 5
Essay Scores: Descriptive Statistics

Mode of Discourse	Grade 3			Grade 4			Grade 5		
	N	M	SD	N	M	SD	N	M	SD
Narrative	100	3.8	1.40	112	5.0	1.75	100	6.1	1.64
Explanation	66	4.1	1.88	101	4.2	1.78	100	4.9	1.75
Description	100	3.6	.99	100	4.5	1.14	100	5.1	1.52
Informative Report	101	3.4	1.37	100	3.9	1.69	100	4.6	1.84
Persuasive Essay	100	4.2	1.90	101	5.3	1.70	72	5.4	1.48

Mode of Discourse	Grade 6			Grade 7			Grade 8		
	N	M	SD	N	M	SD	N	M	SD
Narrative	103	4.7	1.67	101	4.9	1.82	104	5.2	1.81
Explanation	100	3.8	1.60	67	4.6	1.69	95	4.8	2.15
Description	100	4.5	1.49	98	4.6	1.53	42	5.0	2.10
Informative Report	100	4.2	1.60	72	4.8	1.82	100	4.9	1.64
Persuasive Essay	100	4.2	1.68	101	4.8	1.69	99	4.9	1.54

Table 6
Differences between Means: Results of t-tests

Mode of Discourse	Grade 3 M	Grade 3 SD	Grade 5 M	Grade 5 SD	t	Grade 6 M	Grade 6 SD	Grade 8 M	Grade 8 SD	t
Narrative	3.8	1.40	6.1	1.64	10.71***	4.7	1.67	5.2	1.81	2.32*
Explanation	4.1	1.88	4.9	1.75	2.67**	3.8	1.60	4.8	2.15	3.55***
Description	3.6	.99	5.1	1.52	8.42***	4.5	1.49	5.0	2.09	1.43
Informative Report	3.4	1.37	4.6	1.84	5.31***	4.2	1.60	4.9	1.63	3.19**
Persuasive Essay	4.2	1.90	5.4	1.48	4.69***	4.2	1.68	4.9	1.54	3.41***

*p < .05

**p < .01

***p < .001

Table 7
Correlations between Essay Scores and Scores on Objective Measures: Narrative Prompts

	Grade						Average Within-Grade Correlations
	3	4	5	6	7	8	
ITBS							
N	77	111	93	100	35	100	516
Vocabulary	.50	.35	.42	.59	.37	.39	.45
Reading	.65	.41	.48	.61	.61	.42	.52
Spelling	.50	.56	.58	.60	.37	.32	.51
Capitalization	.59	.46	.53	.55	.43	.44	.51
Punctuation	.52	.47	.50	.51	.53	.41	.48
Usage & Expression	.48	.36	.48	.58	.50	.36	.46
Language Total	.60	.52	.57	.63	.56	.46	.56
Visual Materials	.59	.24	.42	.39	.43	.34	.39
Reference Materials	.51	.41	.42	.48	.43	.37	.43
Work-Study Total	.58	.34	.45	.46	.47	.38	.44
Math Concepts	.35	.33	.42	.49	.44	.35	.39
Math Problem Solving	.32	.39	.35	.53	.38	.34	.39
Math Computation	.46	.45	.46	.45	.46	.16	.40
Math Total	.43	.42	.47	.55	.48	.32	.44
Composite	.62	.45	.51	.63	.61	.45	.54
CogAT							
N	85	93	91	94	95	93	551
Verbal	.52	.40	.47	.50	.59	.44	.49
Quantitative	.36	.29	.38	.48	.41	.41	.39
Nonverbal	.36	.21	.31	.31	.41	.33	.32

Table 8
Correlations between Essay Scores and Scores on Objective Measures: Explanation Prompts

	Grade						Average Within-Grade Correlations
	3	4	5	6	7	8	
ITBS							
N	66	93	84	64	66	85	458
Vocabulary	.41	.38	.49	.48	.52	.51	.47
Reading	.45	.40	.53	.54	.50	.51	.49
Spelling	.51	.35	.62	.52	.43	.38	.47
Capitalization	.42	.39	.54	.59	.54	.50	.50
Punctuation	.47	.34	.55	.61	.47	.50	.49
Usage & Expression	.54	.36	.62	.55	.64	.61	.55
Language Total	.55	.41	.65	.64	.59	.56	.57
Visual Materials	.37	.35	.50	.48	.46	.55	.46
Reference Materials	.40	.43	.49	.49	.60	.61	.51
Work-Study Total	.42	.42	.53	.50	.58	.61	.51
Math Concepts	.36	.46	.54	.42	.52	.58	.49
Math Problem Solving	.36	.42	.44	.40	.57	.52	.46
Math Computation	.48	.30	.46	.53	.28	.38	.40
Math Total	.44	.46	.53	.51	.54	.57	.51
Composite	.52	.45	.60	.59	.63	.60	.56
CogAT							
N	55	94	89	96	59	87	480
Verbal	.49	.38	.52	.56	.42	.50	.48
Quantitative	.45	.40	.44	.46	.38	.49	.44
Nonverbal	.47	.37	.43	.48	.24	.42	.41

Table 9
Correlations between Essay Scores and Scores on Objective Measures:
Description Prompts

	Grade						Average Within-Grade Correlations
	3	4	5	6	7	8	
ITBS							
N	98	98	81	67	98	41	483
Vocabulary	.47	.36	.56	.31	.36	.58	.43
Reading	.50	.50	.61	.49	.32	.54	.49
Spelling	.59	.46	.62	.44	.12	.52	.46
Capitalization	.51	.50	.58	.51	.22	.55	.47
Punctuation	.47	.49	.54	.49	.14	.57	.44
Usage & Expression	.51	.37	.57	.43	.15	.58	.42
Language Total	.60	.53	.64	.52	.19	.60	.51
Visual Materials	.51	.45	.57	.28	.29	.48	.47
Reference Materials	.61	.47	.49	.46	.19	.56	.46
Work-Study Total	.61	.48	.56	.42	.26	.57	.48
Math Concepts	.48	.34	.61	.34	.28	.52	.42
Math Problem Solving	.46	.32	.57	.31	.39	.55	.43
Math Computation	.42	.45	.47	.42	.23	.51	.41
Math Total	.52	.42	.62	.42	.34	.56	.48
Composite	.60	.53	.65	.48	.34	.62	.53
CogAT							
N	90	88	94	98	95	34	499
Verbal	.61	.42	.51	.39	.37	.53	.47
Quantitative	.42	.41	.48	.32	.31	.39	.39
Nonverbal	.35	.45	.43	.26	.30	.55	.37

Table 10
Correlations between Essay Scores and Scores on Objective Measures: Informative Reports

	Grade						Average Within-Grade Correlations
	3	4	5	6	7	8	
ITBS							
N	99	79	88	96	72	70	504
Vocabulary	.14	.38	.38	.47	.45	.29	.35
Reading	.30	.41	.39	.54	.52	.27	.41
Spelling	.08	.25	.51	.51	.45	.17	.34
Capitalization	.21	.36	.39	.60	.46	.21	.39
Punctuation	.17	.39	.36	.56	.47	.17	.37
Usage & Expression	.14	.39	.43	.48	.55	.24	.38
Language Total	.18	.40	.47	.60	.55	.24	.42
Visual Materials	.14	.44	.32	.57	.34	.07	.33
Reference Materials	.14	.47	.35	.64	.55	.20	.41
Work-Study Total	.16	.48	.35	.63	.49	.14	.39
Math Concepts	.09	.34	.35	.52	.43	-.09	.31
Math Problem Solving	.13	.41	.30	.53	.34	.20	.33
Math Computation	.15	.30	.50	.54	.56	.03	.37
Math Total	.15	.40	.42	.56	.49	.06	.36
Composite	.25	.46	.43	.60	.57	.24	.44
CogAT							
N	94	90	92	90	63	94	523
Verbal	.28	.46	.44	.48	.39	.42	.41
Quantitative	.09	.36	.41	.54	.38	.17	.33
Nonverbal	.15	.32	.35	.33	.15	.07	.23

Table 11
Correlations between Essay Scores and Scores on Objective Measures:
Persuasive Essays

	Grade						Average Within-Grade Correlations
	3	4	5	6	7	8	
ITBS							
N	68	101	72	100	99	96	536
Vocabulary	.40	.32	.26	.50	.24	.12	.31
Reading	.53	.37	.29	.57	.38	.03	.37
Spelling	.35	.35	.18	.44	.34	.20	.32
Capitalization	.35	.42	.18	.53	.32	.08	.33
Punctuation	.33	.32	.31	.55	.38	.14	.35
Usage & Expression	.30	.37	.38	.58	.30	.21	.37
Language Total	.38	.41	.30	.60	.38	.18	.39
Visual Materials	.37	.35	.33	.48	.34	0	.31
Reference Materials	.33	.39	.26	.54	.39	.01	.33
Work-Study Total	.38	.40	.31	.56	.40	0	.35
Math Concepts	.36	.42	.22	.44	.23	.17	.31
Math Problem Solving	.30	.34	.28	.56	.28	.04	.31
Math Computation	.18	.21	.25	.55	.37	-.02	.28
Math Total	.34	.37	.29	.58	.32	.07	.34
Composite	.48	.42	.33	.62	.40	.10	.40
CogAT							
N	95	96	44	93	92	90	510
Verbal	.40	.43	.33	.54	.30	.17	.37
Quantitative	.46	.37	.24	.50	.14	.15	.33
Nonverbal	.31	.32	.26	.50	.10	.15	.28

Table 12
Summary of Correlations between Essay Scores and Scores on Objective Measures:
Average Within-Grade Correlations

	Narrative	Explanation	Description	Informative Report	Persuasive Essay	Average
ITBS						
Vocabulary	.45	.47	.43	.35	.31	.40
Reading	.52	.49	.49	.41	.37	.46
Spelling	.51	.47	.46	.34	.32	.42
Capitalization	.51	.50	.47	.39	.33	.44
Punctuation	.48	.49	.44	.37	.35	.43
Usage & Expression	.46	.55	.42	.38	.37	.44
Language Total	.56	.57	.51	.42	.39	.49
Visual Materials	.39	.46	.47	.33	.31	.39
Reference Materials	.43	.51	.46	.41	.33	.43
Work-Study Total	.44	.51	.48	.39	.35	.44
Math Concepts	.39	.49	.42	.31	.31	.39
Math Problem Solving	.39	.46	.43	.33	.31	.38
Math Computation	.40	.40	.41	.37	.28	.37
Math Total	.44	.51	.48	.36	.34	.43
Composite	.54	.56	.53	.44	.40	.49
CogAT						
Verbal	.49	.48	.47	.41	.37	.45
Quantitative	.39	.44	.39	.33	.33	.37
Nonverbal	.32	.41	.37	.23	.28	.32

Table 13
Disattenuated Correlations between Essay Scores and *ITBS* Language Total and *CogAT* Verbal Scores

Mode of Discourse	*ITBS* Language Total	*CogAT* Verbal
Narrative	.69	.61
Explanation	.83	.70
Description	.85	.79
Informative Report	.68	.67
Persuasive Essay	.66	.63